You cannot scan a healthcare related newspaper, newsfeed, magazine or website these days without seeing a reference to predictive analytics. That it continues to show up so frequently is a tribute to its growing importance and the fact that talking about it is still “sexy.” But understanding why predictive analytics are coming to the forefront is only a small part of the story. The real story is being able to see how using predictive analytics in your day to day work will help you better understand your business and help you make better business decisions by applying scientific and statistical knowledge and techniques to your data without having to be a data scientist or an expert in science or statistics.

A predictive model is not magic and it is not rocket science. It is actually a mathematical representation of reality that offers a view based on science and statistics that uses data that is currently available to predict the risk of an adverse event (or a preferred event) at some point in the future. The models, and there are many, look for patterns in the data that may not always be evident and which seem linked to an outcome of some sort. Depending on the data and the model it can do so in either a supervised learning fashion, where the model describes a relationship between a set of independent attributes and a dependent attribute or in an unsupervised learning fashion where the model, itself, will find relationships in the data without reference to independent or dependent variables.

One of the very cool things about some of the newer models such as what is available in the Predixion Suite is that they are pretty much agnostic as to the source and the type of data that can be factored into the analysis. The only real restrictions are:

1) Relevant and consistent input data are needed
2) The outcome you are trying to predict must be measurable, and
3) A way to relate the two mathematically must exist

Bullet point number three may seem somewhat self-evident but it explains why a predictive model cannot predict statistically random events such as trauma or pregnancy.

This whitepaper is being written in support of the Predixion Suite of software from Predixion Software, which is a state of the art predictive analytics platform with an interface that is both intuitive and extremely easy to use. **So why bother with reading something called practical predictive analytics for healthcare 101?** Because doing so and internalizing some of the understanding of what predictive models are; how they work; what they can do (and what they cannot do) and how to interpret their results is going to make the Predixion Suite that much more relevant and valuable to you as a practical resource that you can use in your daily work. Its intent is not only to give you an understanding of the models and the tool, but a framework for how to think about structuring your thoughts and data to facilitate the analysis of your data and make predictive analytics a powerful and useful tool in your toolbox.

There are currently well over 100 commercially available predictive models. All use a variety of machine learning techniques, mathematical algorithms, the application of rule sets and other methodologies to accomplish their predictions. And, while Predixion uses mostly statistically based algorithms as the core of its engine, the general techniques and processes we are going to discuss are relevant to all models.
Predictive models are also really not all that new, although healthcare as an industry has been somewhat later “to the party” than other industries. There are published articles going back as far as 1941\(^1\) that have shown how the use of mathematics and computers have consistently been shown to be superior to a panel of experts in actually predicting things directly related to the expertise of the panel members. But at the current time predictive analytics and predictive models and their usage are really starting to explode due to the maturing of multiple technologies coupled with falling prices for very powerful computing resources and data storage, as the following graphic demonstrates.

---

\(^1\) reference article is “50 Years of Successful Predictive Modeling Should be Enough: Lessons for the Philosophy of Science”: Trout, JD and Bishop, M; *Philosophy of Science 68* (Proceedings): S197-S208
Let us spend a bit more time talking about each of the seven (7) steps. Following this approach will give you a general framework for creating your own queries and assure that your likelihood of success will be higher. Much of this may seem somewhat self-evident but you might be surprised how often many of these steps are either skipped over or minimized in order to “get going” on the model and derive a result. Allowing oneself to “jump to getting an answer” can create an output that may either not really answer the question you had to begin with or worse, might actually mislead you. All of us do have a tendency to simply believe what a computer screen tells us. So let’s review the steps.

**Step 1 is defining the problem.** Put simply, *what question(s) are you trying to answer?* Once you understand that you need to then think about what data is available to you to answer the question:

- Is the data directly related to the question?
- If it is not, can you create a proxy relationship to be able to link it?
- Is the data you need even available within the enterprise or elsewhere?

As part of step 1, you also need to specify the inputs and outputs of the model you are going to build as these may change as you change and tweak the model. Finally, don’t forget the most commonly forgotten piece of any new initiative. Determine, up front, how you are going to measure the results.

- What measure of accuracy are you going to use? Is that level of accuracy good enough for the business?
- How will you benchmark the results?
- What criteria are you going to use to determine success or failure?

**Step 2 is more rote and technical – process the data.** Collect the data (more is always better in this analyst’s mind but more does not always mean easier or better results). In general, more recent data is better and the data need to be consistent. Don’t skimp on cleaning the data. While this may end up taking the most time, it is critical and erroneous data will create erroneous results. Transforming the data is also worth the time and effort to improve the modeling process including such things as:

- Converting non-numerical data to numeric (or vice versa)
- Standardizing thing such as coding, definitions, costs, combining variables, etc.

Predixion gives you tools, built into the software, to help you analyze, clean, and classify your data on the front end.

**Step 3 is running the initial model.** Part of step 3 is to split the dataset into a test dataset and a validation dataset. If you really want to be able to test the accuracy of the model once it has been built, you need to do this. The software will walk you through this and will do this for you by default, holding back 30% of the total data for validation testing, although it does allow you to override the default values. This is also the step whereby you will choose the method or methods by which you want to build the model and process the data. As you become more familiar with predictive modeling and with your own data you will find that certain types of data and certain types of analyses or problems do lend themselves more or less to certain types of modeling. But if you are just starting out, you can use the software to guide you in the choice of a model or simply choose to run all the models against your data. Once done, run the model and move on to step 4.
Step 4 is to evaluate the initial results. Are the results acceptable? Are they what you were expecting to see? Do you understand the results? Do they answer the question you are trying to answer? If the answer is yes, then move on to the next step. If the answer is no, then consider the following:

- Try using different algorithms/models
- Try using different data elements or repackaging what you have
- Consider collecting more or different data
- Consider redefining the problem, changing the question and the means to an answer as you better understand your data and your environment

Part of the learning process may be to try and not “boil the ocean” with your models. Think about setting up the model to run on a number of scenarios starting from the simple and most straightforward and then progressing to more and more complex.

Step 5 is to select the final model. Don’t be afraid to try a number of different models and then when you are satisfied with the results, choose the best one. We will talk about means of assessing the accuracy of your model in a bit. For now, choose the final model and consider whether you want to re-run the entire dataset against the selected model and re-examine the results.

Step 6 is to test the final model. This is another one of those things that often seems not to get done. It is important to test the final model and the only way to do so is to take the selected model and run it against a second, unrelated dataset (e.g. - the validation dataset or the portion of the dataset that was held back for this purpose) and assess the results. Do not tweak or change the model in any way at this point as it will invalidate any comparison to the initial results. If the results are similar and you are satisfied with them you can move on to the final step. If you are not, then go back (to step 3) to reassessing the model and the data, make any necessary or desired changes and try re-running the model again.

Step 7 is to apply the model and run the prediction. There are actually two parts to this. One is done if you want to refine the model then you can use the output from the model to determine next steps and potential intervention or changes. Continue to test the model as much or as often as needed. But when you are satisfied, please do two things:

1) Run the necessary measures to test the final accuracy of the model (see our next discussion points)
2) Take the output from the model and turn it back into language and output for the business that answers the initial question you set out to answer and makes it useful/usable for them

Not doing the second part of this will make the exercise you just went through interesting but academic. The key is to use the model and your special knowledge to interpret the results back to the business so they can use and leverage the results.

Now that the model has been run, you have to evaluate the model and make a determination on how accurate the model is. In order to do so you have to understand some things about predictive models and, unfortunately, you need to refresh your understanding of some statistics terminology and techniques. Now, for those of us whose response is “oh my God I thought I was done with this stuff”
there actually is good news. The Predixion tool will run the accuracy checks for you and help you choose and understand the results. But spending just a few minutes right here right now will make the process that much easier.

We’ve already noted that a predictive model cannot predict random events. Since we are restricting this whitepaper to the use of predictive modeling in the healthcare arena there are also a couple of other “givens” that must also be pointed out. In healthcare most all models are built at the individual person (member or patient) level and then aggregated up for looking at the results at a population or sub-population level. The reported accuracy of a model, however, is always reported at the population level and that does not mean that you can apply the same level of accuracy to an individual. You can’t. All models do a much better job predicting at the population, rather than the individual level. Sorry, but that is just the way it works.

In addition, all of the models tend to over-predict at the lower end and under-predict at the upper end so if what you are trying to predict occurs naturally in either of these tails or are actually in the range of what might be called outliers than the predictive power of the model will not be as good.

In order to evaluate a model you ideally need to be able to evaluate large populations with multiple variables some of which may act independently and some of which may interact in an interdependent manner. The technical term for that type of analysis is multivariate analysis. Though certainly not a totally inclusive listing, some knowledge of the following terms will help you understand the area of predictive modeling better and make it easier for you to judge the value and the accuracy of models that you create. Those terms are:

- Sensitivity and Specificity
- True and False Positives and Negatives
- Predictive Value and Positive Predictive Value
- Parametric vs. Non-Parametric
- Correlation Coefficient (R) and R-Squared
- Regression
- ROC curve (aka area under the curve)

My goal here is not to create an elementary book of statistics but simply give you enough knowledge and ammunition to be able to assess how good (or bad) a given model is. For most of you this will be something you have seen before. For some it may seem elementary. For any of you who find this totally foreign, it is certainly possible to use the Predixion tools and build good models without this level of knowledge but I would encourage you to spend a bit of time to gain some understanding of basic statistics. There are lots of good sources of knowledge out there, many are free, and the time spent will make your work easier and more productive.

Let me cheat a bit by using the following slides taken from the same presentations referenced earlier. The comments to the right are new and relevant to this discussion. While these use some examples it should give you the definitions required.
Positive predictive value is a very useful technique to be familiar with as it is a way of judging the precision of certain types of models. As noted, it is a way of reflecting the performance of a diagnostic method or model by being a direct reflection of the probability that a positive result actually reflects the underlying condition or set of conditions that are being tested for. It is particularly useful in Bayesian Analysis which is one of the methods available to you within the Predixion Suite. However, you should be aware that its value is dependent on the prevalence of the outcome of interest which is something you may not always know on a particular set of data or target population. It is also influenced by the threshold that you establish for calculating the probability as demonstrated in the following slide:
I find it helpful to think about the type of data that I am looking at when I am in the process of setting up an analysis and choosing a model for predictive purposes. This introduces the terms of **parametric** and **non-parametric** data. In the simplest of explanations, parametric data is normally distributed (such as a normal distribution or “bell” curve for any number of things and non-parametric data is not normally distributed.

Just as you would expect, increasing the threshold of your cut point raises the number of true positives (specificity) but lowers the % of true positives you can predict (sensitivity)

Parametric data is usually (but not always) reasonably easy to recognize. When you are looking at a new dataset, though, you may not be sure. If you are not sure, then you may want to consider the data non-parametric.

Right around now you should be asking yourself “why does any of this matter?” It does matter because in order to create the best model and in order to best understand the outcome of your model you need to choose the right test or model for the data that you have. There are, as noted, hundreds of models and hundreds of tests that can be applied. Once again, Predixion does much of the work for you by giving you tools to clean and explore your data and to guide you in choosing the best path. But the more you understand some of this, the easier your model building is going to be and the better your results are going to be.
So why wouldn’t you just choose non-parametric tests/models all of the time? You can and sometimes you may need to but they are also less powerful than parametric tests. Simply put, because they do not make any assumptions about the distribution of the data they have less information that they can utilize to determine significance. As a result, it is harder to show statistical significance using non-parametric methods. Or, put another way, you may have to be able to demonstrate stronger results in order to classify the results as significant or the prediction as showing a higher level of probability.

Regression is one of the most commonly used methods for building predictive models and, in my opinion, is a particularly powerful way of doing so. Simply put, regression is a technique of fitting a simple equation to real data points. Linear regression (fitting the data points to a line) is the most common type but there are others such as multi-linear regression and logistic regression (one of my favorites). It is a mathematical way of assessing the impact and contribution of diverse/disparate variables on a process or outcome. Typically, linear regression is used for variables that are continuous and logistic regression is used for variables that are binomial (yes or no; off or on; present or not present). Understanding this can also help you modify or change your data in such a way as to make the data better fit the type of model you want to run.

R-squared is a term that most of you have either heard of or are familiar with. R is actually what is called a correlation coefficient and it measures the strength of a linear relationship between two things. R-squared is what it says. It is the square of the correlation coefficient, i.e., the total squared error that can be explained by the model and is an excellent and common way of expressing the relative predictive power of a model. It is a descriptive measure and it is always reported as a number between zero (0) and one (1). R-squared explains the variance of the model or the deviation from what was expected. Put another way, it tells you how much your prediction is improved by applying linear regression to your data compared to not doing so. No model can ever achieve a score of 1 and the results are typically expressed as a fraction. So, an R-Squared of 0.35 (which is a commonly seen number for good prospective medical models) means that the model can explain 35% of the variance that is seen in the data. Comparing the result to some other known methods (e.g. – age-sex as a predictor yields an R-Squared in the range of 0.05) can give you a good sense of how good or bad the model might be. But you have to remember that the R-squared really only makes sense when there is a linear relationship between the data elements which, by definition implies parametric data.

I am going to close this section with a brief discussion of my favorite way to test a model, the ROC curve, i.e., the area under the curve. ROC actually stands for receiver operating characteristics and it was developed in the 1950’s for understanding radio signals that were being contaminated by noise. But they are also very useful as an aid to decision making in the healthcare space. They work well for data that may not conform to a normal distribution and they are focused on the use of a threshold to determine whether something is positive or negative. And, that threshold can be changed and tested to determine the best cut point.
A simple example may help demonstrate this:

When the threshold is very high, there will be almost no false positives... but we won’t really identify many true positives either. As you move the test threshold towards a more reasonable, lower value - number of true positives increases (dramatically at first, so the ROC curve moves steeply up) - Eventually you reach a region where there is a remarkable increase in false positives

Similar to the concept of R-squared, you will never reach the ideal model state but as you do this, you will quickly gain a sense of “goodness of fit” for your model. The generic example above is, in my opinion, an example of a pretty darn good model where the blue line indicates no model, the yellow is an ideal model, and the purple represents the model you have built.

To avoid devolving into too much of a technical discussion I am going to stop here. The examples and the documentation that are included with the Predixion Suite will help you understand this and the other types of model testing that are available to you within the tool as you build your models. And, as I have repeatedly said, it will help guide you through the selection process.

Hopefully this white paper has given you a refresher course on some things and stimulated your thought processes and raised some questions in your mind on others. If so, then I have succeeded in what I set out to do.

Predixion’s predictive analytics technology, in partnership with Avanade’s industry expertise, makes it easy for you and you really do not need to know or understand any of this to be able to actually use the solution and run predictive models to answer questions for your business. That is part of the beauty and it is a testament to the knowledge and the abilities of Predixion’s development team. It is not easy to take something that is actually quite technical and complex and make it into an easy to use, easy to understand solution with a very intuitive interface that does not require a large learning curve to make it useful. But that is exactly what Predixion, in partnership with Avanade, has done.
But that very simplicity also masks the real power of this solution and I hope that I can engage you a bit to strive to unleash that power as you get to know and use the tool. By starting to understand what is going on “underneath the hood” and investing a little bit of time in the “stuff” we have been discussing you will start to gain a better understanding of your data and how you can approach (and change/modify) your data to make it more amenable to a predictive model. In the end, you will know your data better; you will be able to phrase and frame the business questions you need to answer better; you will be able to choose or create the type of data you need to answer those questions better; and in the process you will have made yourself a better analyst.

And, by the way, you will also have solved some problems for your business and made yourself a more valuable asset to that business at the same time.

This is a journey and it is a fun one that is only going to increase in its need in your day to day business and its importance over time. I wish you well on that journey. Your choice of Predixion in partnership with Avanade has set you on the right course and given you an edge right out of the box.

Bon Voyage and Bon Chance!

**About the Author**

Dr. Eisenberg is an experienced physician executive and medical informaticist with broad experience in medicine, healthcare administration, computer systems (both hardware and software), development, use and interpretation of clinical data and business administration. He is an independent consultant in healthcare analytics and has held senior management positions at a number of organizations including two Blue plans and two disease management companies as well as the position of Chief Science Officer at UnitedHealthcare. He started working with the Predixion Suite in November of 2011 and liked it so much that he ended up joining the company as a member of the Tiger Team. He is currently engaged on contract as the Chief Medical Scientist for a startup telemedicine care delivery organization (LifeView) in the Twin Cities focusing on the very sick and very high cost top 1% of the population as well as working with Predixion and its customers in the implementation and usage of the tools and testing, expanding, and refining the tools.

He also practiced medicine for 22 years, has taught medical students and residents including a stint as a residency program director and assistant dean and has been involved as a beta tester and/or co-development partner in a number of commercially available software packages over the past 20 years. He is a published author, a contributing journal editor, and experienced public speaker. His research experience and interests are in the areas of predictive modeling, medical quality, performance assessment and benchmarking.

**About the Slides**

All slides were taken from presentations given by the author. Those that are not referenced represent the author’s original work. Slides that contain references represent the adapted public work of others. As it may be a bit difficult to see the references within the whitepaper, they are re-listed here.

Slide “Sensitivity vs Predictive Value & Threshold source” = “Using Predictive Modeling to Target Interventions” by Barry P. Chaiken, MD, MPH (CMO at ABQAURP-PSOS)
Slides “Understanding ROC” and “ROC Curves” source = jo@www.anaesthetist.com
About Predixion Software
Predixion Software develops predictive analytics technology that integrates with leading business intelligence platforms, clinical systems and business applications. Predixion offers an easy-to-deploy self-service predictive analytics solution that allows customers to unlock deep insights within their data. Predixion’s healthcare solutions group is committed to finding new and innovative ways to apply our technology to the challenges created by the rapidly evolving healthcare economy. As the industry moves towards wellness based payment structures Predixion will continue to help payers and providers identify at risk patients before they suffer catastrophic and costly events.

Solutions for Healthcare Providers:
- Managing Preventable Readmissions
- Length of Stay Estimation
- Detecting At Risk Populations
- Hospital Acquired Condition Prevention
- Identifying Overpayment
- Medication Adherence
- Membership Attrition Management
- Fraud / Error Detection

Predixion Software, founded in 2009, is headquartered in Orange County, California with development offices in Redmond, Washington.

Predixion Software
31910 Del Obispo #120
San Juan Capiistrano, CA 92675 USA
+1.949.373.4900
www.predixionsoftware.com

About Avanade
Avanade provides business technology solutions and managed services that connect insight, innovation and expertise in Microsoft® technologies to help customers realize results. Our people have helped thousands of organizations in all industries improve business agility, employee productivity, and customer loyalty. Avanade combines the collective business, technical and industry expertise of its worldwide network of experts with the rigor of an industrialized delivery model to provide high quality solutions using proven and emerging technologies with flexible deployment models—on premises, cloud-based or outsourced. Avanade, which is majority owned by Accenture, was founded in 2000 by Accenture LLP and Microsoft Corporation and has 17,000 professionals in more than 20 countries. Additional information can be found at www.avanade.com.

Avanade and the Avanade logo are registered trademarks or trademarks of Avanade Inc. Other product, service, or company names mentioned herein are the trademarks or registered trademarks of their respective owners.

Avanade
North America
Seattle
Phone +1 206 239 5600